To miss-attend is to misalign! Residual Self-Attentive Feature Alignment for Adapting Object Detectors
(Supplementary material)

Anonymous WACV submission

Paper ID 1217

In the supplementary materials, we present comparative detection results and visualization analysis for our framework **(ILLUME)** and the state-of-the-art [4].

1. Additional Detection Examples

We present additional comparative detection results in Figure S1 for the state-of-the-art and ILLUME. The detection results are shown on the target domain (Foggy Cityscapes) [6] for the weather adaptation task. We can see the improved detection performance using our framework. The instances are successfully detected even in extreme weather conditions (foggy weather) as compared to state-of-the-art where most of the instances remain undetected due to domain gap between source (Cityscapes) [1] and target data (Foggy Cityscapes) [6].

In Figure S1, we can see instances such as train and person that are successfully detected using ILLUME. As can be seen in the first row of the figure, where the train instance (right side of image) is correctly detected using ILLUME, while state-of-the-art misses it. Also, in the third row, we can see that state-of-the-art fails to detect multiple bicycle instances as compared to our detection results where multiple bicycles (on the left of the image) are detected successfully. Similarly in other comparative examples in the figure, we can see that most instances like persons or cars remain undetected in the state-of-the-art results mostly due to domain gap (foggy weather); in contrast, they are correctly detected using ILLUME. This proves the effectiveness of our method (ILLUME) to improve detection performance as it focuses on enhancing important instances in the images.

2. Visualization Analysis

In Figure S2, we present a detailed qualitative visualization analysis of the enhanced features using our method (ILLUME) compared with state-of-the-art. We use target samples from the Foggy Cityscapes dataset [6] (weather adaptation task) for this analysis. These features are the transformed features, which are the output of the detection backbone network. A clear comparison can be seen between state-of-the-art and ILLUME. Our method correctly highlights important instances in the image, like car or bike which are missed by [4]. As seen in the second and third rows, many instances are missed in the feature maps of state-of-the-art, or inaccurately highlighted or enhanced, as seen in the first row. In contrast, our methods enhances the objects of interest as required, and hence does not miss them. The enhanced features depict the effectiveness of our method to transform features such that only important instance features would be considered by Faster R-CNN [5] to learn domain-invariant features essential for alignment. In our paper, we also perform similar visualization analysis for two different domain adaptation tasks: (1) Weather Adaptation (Cityscapes [1] to Foggy Cityscapes [6]); and (2) Dissimilar Domain Adaptation (Pascal VOC [2] to Clipart [3]), in Sections 4.4 and Figure 3. It is worth noticing that visualizations for both source (Cityscapes) and target (Foggy Cityscapes) instances are similar irrespective of the domain gap, as shown in Figure 3 of our main paper. These results also corroborate the claim of our method's effectiveness in aligning the instances well with improved enhancement of the feature maps – thereby improving detection performance.

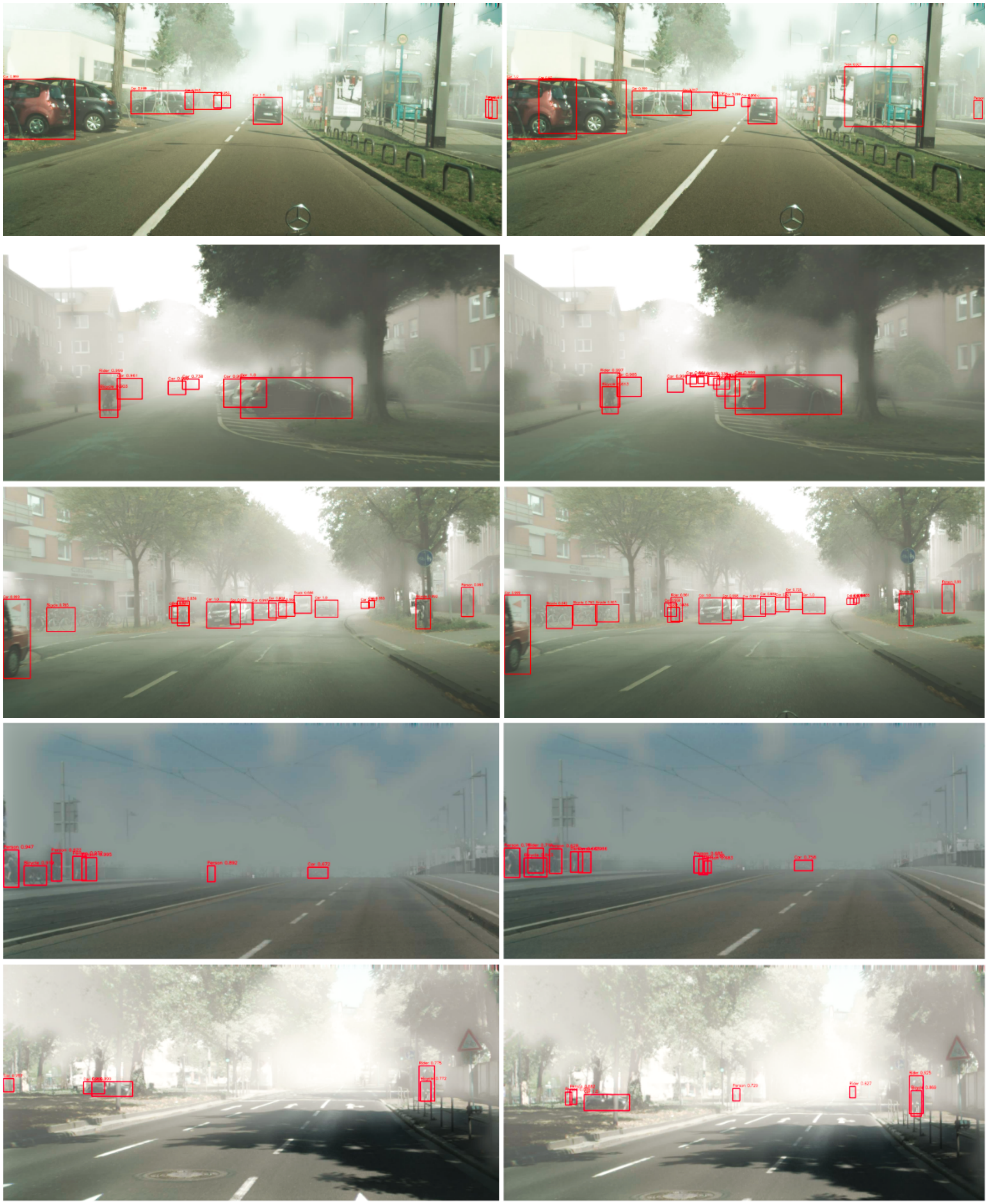
References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 1

[3] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-Domain Weakly-Supervised Object De-

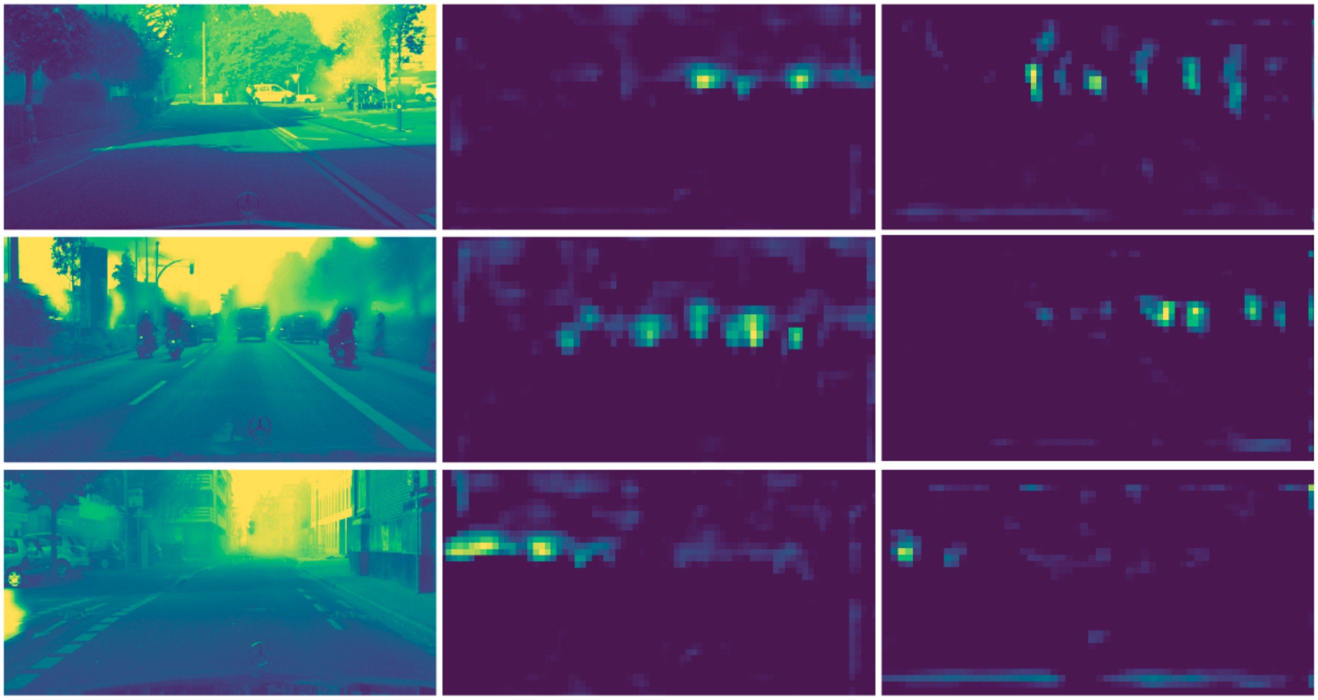
WACV
#1217



a) State-of-the-art

b) Ours (Illume)

Figure S1. **Comparative Detection Results:** Improved detection performance can be seen using our method (ILLUME), compared to state-of-the-art [4] that fails to detect instances like train in first row, bicycles in third, and persons or cars in other.



a) Target samples b) Ours (Illume) c) State-of-the-art

Figure S2. **Visualization Analysis :** Visualizations of transformed features using our ILLUME method that enhances important instances successfully. On the other hand, the state-of-the-art [4] inaccurately highlights instances as seen in the top row, as well as misses important instances as seen in the second and third rows.

tection Through Progressive Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[4] Shuai Li, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Category dictionary guided unsupervised domain adaptation for object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1949–1957, 2021. 1, 2, 3

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 1

[6] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic Foggy Scene Understanding with Synthetic Data. *International Journal of Computer Vision*, 126(9):973–992, 2018. 1